MATH343 LEC 13

Recall $\vec{b} = (X^T X)^{-1} X^T \vec{y}$ was the result from $\underset{\vec{w} \in \mathbb{R}^{p+1}}{\text{argmin}} \|\vec{y} - X\vec{w}\|^2$

This is not one of the estimation-generating procedures we talked about in 341.

Consider $\vec{\mathcal{E}} \sim N_n(\vec{0}_n, \sigma^2 I_n)$. What is the MLE for $\vec{\beta}$?

our assumption

$$\Rightarrow \vec{Y} \sim N_n(X\vec{\beta}, \sigma^2 I_n) = f(\vec{y}; \beta, \sigma^2, X) \Rightarrow$$

$$\mathcal{L}(\vec{\beta}, \sigma^2; \vec{y}, X) = N_n(X\beta, \sigma^2 I_n) = \frac{1}{(2\pi)^{n/2}\underbrace{\sqrt{\det[\sigma^2 I_n]}}_{(\sigma^2)^{-\frac{n}{2}}}} e^{-\frac{1}{2}(\vec{y}-X\vec{\beta})^T \underbrace{(\sigma^2 I_n)^{-1}}_{\frac{1}{\sigma^2}I_n}(\vec{y}-X\vec{\beta})}$$

$$\ell(\vec{\beta}, \sigma^2; \vec{y}, X) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\|\vec{y}-X\vec{\beta}\|^2$$

$$\frac{\partial}{\partial\vec{\beta}}[\ell] = \cancel{-\frac{1}{2\sigma^2}}\frac{\partial}{\partial\vec{\beta}}[\|\vec{y}-X\vec{\beta}\|^2] \overset{set}{=} 0 \Rightarrow \frac{\partial}{\partial\vec{\beta}}\left[\sum(\vec{y}_i - \vec{x}_{i\cdot}\vec{\beta})^2\right] = 0$$

This is exactly the same problem solved before! $\Rightarrow \hat{\vec{\beta}}^{MLE} = (X^TX)^{-1}X^T\vec{y} = \vec{b}$,

i.e. the least squares estimate is the MLE for β!

Note: this is the MLE regardless of the value of σ^2!

Analogue: $\bar{X}$ was $\hat{\theta}^{MLE}$ for $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ regardless of value of σ^2!

$$\frac{\partial}{\partial\sigma^2}[\ell] = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\|\vec{y}-X\vec{\beta}\|^2 \overset{set}{=} 0 \Rightarrow -n + \frac{1}{\sigma^2}\|\vec{y}-X\vec{\beta}\| = 0$$

substitute $\vec{\beta} = \hat{\vec{\beta}}^{MLE}$

$$\Rightarrow \hat{\sigma}^{2\,MLE} = \frac{1}{n}\|\vec{y}-X\vec{\beta}\| = \frac{1}{n}\|\vec{y}-X(X^TX)^{-1}X^T\vec{y}\|$$

$$= \frac{1}{n}\|(I-H)\vec{y}\|^2$$

$$= \frac{1}{n}\|\vec{e}\|^2 = \frac{1}{n}\sum e_i^2 = \frac{1}{n}SSE$$

Oh HW!!

Remember in 342w, if $p+1 > n$, OLS fails since X^TX is no longer invertible. But what if we still want inference?

Bayesian shrinkage estimators can help us.

Let's assume a prior on $\vec{\beta}$: $\beta_0, \beta_1, \ldots, \beta_p \overset{iid}{\sim} N(0, \tau^2)$

$$\Rightarrow f(\vec{\beta}) = N_{p+1}\left(\vec{0}_{p+1}, \tau^2 I_{p+1}\right).$$

If $\tau^2 \to \infty$ this prior is the Laplace flat prior for all β_j

And also on σ^2

$$f(\sigma^2) \propto \frac{1}{\sigma^2} \propto \text{InvGamma}(0,0) \text{ and independent of } \vec{\beta}$$

$$f(\vec{\beta}, \sigma^2 \mid X, \vec{y}) \propto f(\vec{y} \mid \vec{\beta}, \sigma^2, X)\, f(\beta, \sigma^2 \mid X)$$

$$= \frac{1}{(2\pi)^{n/2} \underbrace{\sqrt{\det[\sigma^2 I_n]}}_{(\sigma^2)^{-n/2}}} e^{-\frac{1}{2}(\vec{y} - X\vec{\beta})^T (\sigma^2 I_n)^{-1} (\vec{y} - X\vec{\beta})} \cdot$$

$$\frac{1}{(2\pi)^{\frac{p+1}{2}} \sqrt{\det[\tau^2 I_{p+1}]}} e^{-\frac{1}{2}(\vec{\beta} - \vec{0}_{p+1})^T (\tau^2 I_{p+1})^{-1} (\vec{\beta} - \vec{0}_{p+1})} \cdot \frac{1}{\sigma^2}$$

$$\propto (\sigma^2)^{-\frac{n}{2} - 1} e^{-\frac{1}{2\sigma^2}\|\vec{y} - X\vec{\beta}\|^2} e^{-\frac{1}{2\tau^2}\|\vec{\beta}\|^2}$$

Since we care currently about inference for $\vec{\beta}$, let's find the marginal posterior:

$$f(\vec{\beta} \mid X, \vec{y}) = \int_0^\infty f(\vec{\beta}, \sigma^2 \mid X, \vec{y})\, d\sigma^2 \propto e^{-\frac{1}{2\tau^2}\|\vec{\beta}\|^2} \int_0^\infty \underbrace{(\sigma^2)^{-\frac{n}{2} - 1} e^{-\frac{\|\vec{y} - X\vec{\beta}\|^2/2}{\sigma^2}}}_{\text{Kernel for inversegamma}(a,b)} d\sigma^2$$

where $a = \frac{n}{2}$, $b = \|\vec{y} - X\vec{\beta}\|^2/2$

$$= e^{-\frac{1}{2\tau^2}\|\vec{\beta}\|^2} \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\|\vec{y} - X\vec{\beta}\|^2/2\right)^{n/2}} \propto e^{-\frac{1}{2\tau^2}\|\vec{\beta}\|^2} \|\vec{y} - X\vec{\beta}\|^{-n}$$

This is not the result of a known kernel. To do inference, we could Gibbs sample it. But if we just care about estimation, we can calculate the $\hat{\vec{\beta}}^{map}$

$$\hat{\vec{\beta}}^{map} := \underset{\vec{\beta}}{\text{argmax}} \left\{ f(\vec{\beta}, \sigma^2 \mid X, \vec{y}) \right\} = \underset{\vec{\beta}}{\text{argmax}} \left\{ \ln\left(f(\vec{\beta}, \sigma^2 \mid X, \vec{y}) \right) \right\}$$

$$= \underset{\vec{\beta}}{\text{argmax}} \left\{ \overbrace{\left(-\tfrac{n}{2}-1\right) \ln(\sigma^2)}^{\text{constant}} - \frac{1}{2\sigma^2} \|\vec{y} - X\vec{\beta}\|^2 - \frac{1}{2\tau^2} \|\vec{\beta}\|^2 \right\}$$

$$= \underset{\vec{\beta}}{\text{argmax}} \left\{ \underbrace{-\frac{1}{2\sigma^2}}_{\text{constant}} \left(\|\vec{y} - X\vec{\beta}\|^2 + \frac{\sigma^2}{\tau^2} \|\vec{\beta}\|^2 \right) \right\}$$

minimize

$$= \text{argmin} \left\{ + \|\vec{y} - X\vec{\beta}\|^2 + \frac{\sigma^2}{\tau^2} \|\vec{\beta}\|^2 \right\}$$

Now if we just care about an estimation technique a/g 3P2W....

let $\lambda = \frac{\sigma^2}{\tau^2} \geq 0$ $\quad = 0$ if $\tau^2 \to \infty$

$$= \text{argmin} \left\{ SSE + \lambda \|\vec{\beta}\|^2 \right\}$$

"Ridge Regression Estimator"

hyperparameter you choose. If $\lambda = 0$ $\Rightarrow R = OLS$.

If λ is large $\vec{\beta} \approx \vec{0}_{p+1}$ which forces the model to shrink b_j's towards $= 0$ β_j's that are very useful for prediction.

$=$

$\vdots$

see next page

$\vdots$

$$\vec{b}_{ridge} = (X^\top X + \lambda I_{p+1})^{-1} X^\top \vec{y}$$

$\hat{y}_{ridge} = X \vec{b}_{ridge} \in \text{Colsp}[X]$ but it is not the orthogonal projection

This process is known as "regularization" as it's making the $\vec{b}$ estimates more "regular" i.e. simpler, closer to $\vec{0}$

$$= \frac{\partial}{\partial \vec{\beta}} \left[(\vec{y} - X\vec{\beta})^T (\vec{y} - X\vec{\beta}) + \lambda \vec{\beta}^T \vec{\beta} \right]$$

$$= \frac{\partial}{\partial \vec{\beta}} \left[\vec{y}^T \vec{y} - 2\vec{\beta}^T X^T \vec{y} + \vec{\beta}^T X^T X \vec{\beta} + \vec{\beta}^T (\lambda I_{p+1}) \vec{\beta} \right]$$

$$= \frac{\partial}{\partial \vec{\beta}} \left[-2\vec{\beta}^T X^T \vec{y} + \vec{\beta}^T (X^T X + \lambda I_{p+1}) \vec{\beta} \right]$$

$$= -\cancel{2} X^T \vec{y} + \cancel{2} (X^T X + \lambda I_{p+1}) \vec{\beta} \overset{set}{=} 0 \quad \text{to find min.}$$

$$\Rightarrow X^T \vec{y} = (X^T X + \lambda I_{p+1}) \vec{\beta} \Rightarrow \vec{b}_{ridge} = \underbrace{(X^T X + \lambda I_{p+1})}_{} {}^{-1} X^T \vec{y}$$

always invertible if $\lambda > 0$ as it forces full rank.

Recall that OLS was an unbiased estimate:

$$\vec{B} = (X^T X)^{-1} X^T \vec{Y}, \qquad E[\vec{Y}] = E[X\beta + \vec{\mathcal{E}}] = X\beta + E[\vec{\mathcal{E}}] = X\beta + \vec{0}_n = X\vec{\beta}$$

$$E[\vec{B}] = E[\,\cdot\,] = (X^T X)^{-1} X^T E[\vec{Y}] = (X^T X)^{-1} X^T X \vec{\beta} = \vec{\beta}$$

Ridge is biased:

$$E[\vec{B}_{Ridge}] = (X^T X + \lambda I_{p+1})^{-1} X^T X \vec{\beta} \quad \text{as } \lambda \to 0 \Rightarrow \text{Bias} \to 0.$$

But it compensates for this bias in lower variance!

Hence the total MSE is usually lower!